Ask·Me·Why: Model-agnostic Explainability Framework

Jonathan Lai and Hegler Tissot (Advisor)

Drexel University

PROBLEM & MOTIVATION

The Challenge

Current explainability methods require access to model internals. What if we could explain ANY model's predictions using only the data?

The Black Box Challenge

Healthcare: Why was this patient diagnosed with condition X?

Finance: Why was this loan application rejected?

Law: Why was this case flagged as high-risk?

Current Limitations

Method	Limitation
SHAP	Requires model access; Different models → different explanations
LIME	Model-specific; Unstable for mixed data types
XGBoost	Tied to one model; Often global rather than instance-specific

OBJECTIVES

1. Explain ANY model without access using neighborhood analysis - Identify instance patterns & provide complementary insights to SHAP/LIME/XGBoost
2. Enable regulatory compliance for healthcare, finance, law
3. Support mixed data types & scale efficiently
4. Empower domain experts with interpretable explanations
5. Provide model prediction explanations

METHODS

Key Insight

Similar instances in embedding space should behave similarly. When they don't, those differences explain the prediction.

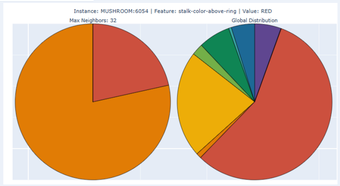

Figure 1: Local vs. Global

Core Innovation

Weighted Cosine Difference (WCD):

$$\mathrm{WCD}(f,x) = \mathrm{CosDiff}(P_{\mathrm{local}}, P_{\mathrm{global}}) \times \mathrm{Uniqueness}(x_f)$$

Where:

P_local = Feature distribution in k-nearest neighbors

P_global = Feature distribution in entire dataset

$$\mathrm{CosDiff} = 1 - \frac{P_{\mathrm{local}} \cdot P_{\mathrm{global}}}{||P_{\mathrm{local}}|| \cdot ||P_{\mathrm{global}}||}$$

$$\mathrm{Uniqueness}_{\mathrm{cat}}(x_f) = 1 - P_{\mathrm{local}}(x_f)$$

$$\mathrm{Uniqueness}_{\mathrm{cont}}(x_f) = \frac{|x_f - \mu_{\mathrm{local}}|}{\sigma_{\mathrm{local}}}$$

Where:

x_f = instance's feature value

μ_local = neighborhood mean

σ_local = neighborhood standard deviation

Framework:

📊 Data→🧠 Embed→🔍Analyze→📦Model→💡Explain

Algorithm Pipeline:

1. Embed instances using unsupervised methods
2. Build KD-Tree index for neighbor search
3. Find k-nearest neighbors in embedding space
4. Compare local vs global feature distributions
5. Weight by instance uniqueness → WCD score
6. Train model on same instances
7. Provide model explanations

Technical Details

Feature Type	Missing Value Handling	Methods
Continuous	Treated as "Missing" bin; excluded from numeric bins	• Adaptive binning (0.25σ, 0.5σ, 1σ, 2σ, 4σ) • Linear interpolation for bin assignment • Distance-weighted aggregation
Categorical	Treated as "Missing" category	• Frequency-based distributions • Distance-weighted counting

Datasets

Dataset	Instances	Features	Task	Characteristics
Mushroom	8,124	22 categorical	Binary	Colors & Structure
Genetic Disorder	22,383	6 cont, 36 cat	Multi-label	Age & lab patterns

Neighborhood Search:

KD-Tree indexing: $O(n \log n)$

k ∈ {16, 32, 64, 128, 256, 512, 1024}

Max radius: 2.0 (Euclidean)

Distance weighting:

$$w_i = \frac{r_{\max} - d_i}{r_{\max}}$$

d_i = distance to neighbor

r_max = maximum radius

Comparison Methods:

SHAP (TreeExplainer) - Shapley value approximation

LIME (Tabular) - Local linear surrogate

XGBoost - Global feature importance

Validation Approach:

2 diverse datasets with different characteristics

Spearman rank correlation for method comparison

Top-k feature agreement analysis

RESULTS

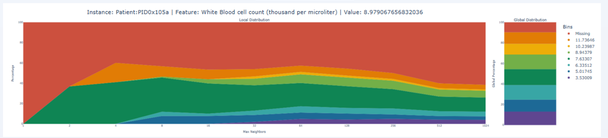

Figure 2. Local vs. Global White-Blood-Cell Distributions

feature	Rank	cos_diff	uniqueness	weighted_cos_diff_normalized
White Blood cell count	1	0.41886	1.931932	0.33363

White Blood Cell count "neighborhood" shifts as we vary the number of nearest neighbors (k). The right-hand bar is the fixed global distribution.

Feature Ranking Comparison:

Table 1: Mushroom Dataset (128 neighbors)

Method Comparison	Spearman ρ	Interpretation
SHAP	0.016	Weak positive
LIME	0.450	Moderate positive
XGBoost	0.186	Weak positive

Table 2: Genetic Disorder (512 neighbors)

Method Comparison	Spearman ρ	Interpretation
SHAP	0.128	Weak positive
LIME	0.121	Weak positive
XGBoost	0.005	Neutral

Model Explanations

1. Mushroom Instance #3900: **(#1 XGBoost: Odor)**

Prediction: EDIBLE ✓ (99.97% confidence)

Top Features (ASK·Me·Why):

1. Gill-color (BROWN)
2. Stalk-color below ring (GRAY)
3. Cap-color (BROWN)

2. Patient PID0x7005:

Prediction: Mitochondrial disorder ✓ (67.9%)

Top Features (ASK·Me·Why):

1. Mother's age (33)
2. Patient age (0)
3. Father's age (52)

CONCLUSIONS

1. Ask·Me·Why produces per-instance explanations **without** model access.
2. **Weighted Cosine Difference** highlights features both distributionally distinct and value-extreme.
3. Complements SHAP, LIME & XGBoost by surfacing **different** but meaningful signals and comparing feature rankings.
4. Scales to mixed data & large datasets via KD-Tree k-NN search.
5. Provides visualization tools for further analysis

REFERENCES

Ribeiro M.T., Singh S., Guestrin C. "Why Should I Trust You? Explaining the Predictions of Any Classifier." *KDD '16*, pp. 1135–1144.

Lundberg S.M., Lee S.-I. "A Unified Approach to Interpreting Model Predictions." *NeurIPS '17*, pp. 4765–4774.

Chen T., Guestrin C. "XGBoost: A Scalable Tree Boosting System." *KDD '16*, pp. 785–794.

Linardatos P., Papastefanopoulos V., Kotsiantis S. "Explainable AI: A Review of Machine Learning Interpretability Methods." *Entropy* 23(1): 18, 2021.